СИСТЕМНЫЙ АНАЛИЗ

И.В. СЕРГИЕНКО, В.В. РЯЗАНОВ, Б.А. БЕЛЕЦКИЙ, А.В. БЫЦЬ, А.М. ГУПАЛ, С.С. РЖЕПЕЦКИЙ

УДК 519.68

МЕТОДЫ ПРЕДСКАЗАНИЯ ПРОСТРАНСТВЕННОЙ СТРУКТУРЫ БЕЛКОВ[1]

Ключевые слова: *распознавание, биофизические фильтры, карты контактов, байесовская процедура, цепь Маркова, фолдинг белка.*

ВВЕДЕНИЕ

На сегодняшний день расшифрованы геномы человека, шимпанзе, мыши, курицы, рыбы Tetraodon и некоторых других животных, нескольких видов растений и грибов, а также свыше тысячи бактерий. Основной вопрос современной молекулярной биологии: какую функцию выполняет определенный ген? Ген — часть молекулы ДНК, которая кодирует белок. Зная нуклеотидную последовательность гена, можно однозначно определить аминокислотную последовательность белка, так как каждая из 20 аминокислот кодируется определенным триплетом нуклеотидов (кодоном). После трансляции последовательности аминокислот из молекулы РНК белок сразу начинает сворачиваться в пространственную конфигурацию. Именно пространственная конфигурация белка определяет его функциональность, поскольку белки в живых организмах взаимодействуют как трехмерные объекты в пространстве. Поэтому в исследованиях белков и их функций придерживаются принципа «последовательность-структура-функциональность» [1]. Это означает, что функциональность белка определяется его пространственной структурой, а пространственная конфигурация задается его аминокислотной последовательностью.

1. ПОСТАНОВКА ЗАДАЧИ

Существует четыре уровня структуры белка:

- первичная — линейная последовательность аминокислотных остатков в молекуле белка;
- вторичная — формирование на линейной последовательности локальных регулярных структур: α-спиралей и β-слоев;
- третичная — расположение элементов вторичной структуры (α-спиралей и β-слоев) в пространстве относительно один другого;
- четвертичная — формирование белкового комплекса из отдельных белков.

Структура белка на каждом уровне оказывает решающее влияние на формирование структуры на следующем уровне, т.е. первичная структура определяет вторичную, вторичная — третичную и т.д.

Первичная структура белка, т.е. его аминокислотная последовательность, находится экспериментальным путем относительно просто. Определение вторичной структуры уже связано с большими сложностями, поскольку требует применения

[1] Работа выполнена в рамках проекта НАН Украины и Российского фонда фундаментальных исследований 2008–2009 гг. при финансовой поддержке Президиума НАН Украины.

дорогих методов рентгеноструктурного анализа и магнитно-ядерного резонанса. Высокая стоимость экспериментального определения структуры белка способствует развитию математических методов ее предсказания. Задача ставится следующим образом: имеется первичная структура белка (т.е. линейная последовательность аминокислот), необходимо определить его третичную структуру, иными словами, отыскать пространственные координаты всех аминокислотных остатков, входящих в белок.

2. МЕТОДЫ МИНИМИЗАЦИИ ЭНЕРГИИ

Подходы к предсказанию структуры белков основаны на термодинамической гипотезе, которая постулирует, что в естественном свернутом состоянии белка свободная энергия системы «белок — растворитель» минимальна. Эта свободная энергия состоит из межмолекулярного взаимодействия самого белка и свободной энергии сольватации (см. ниже).

Исходя из термодинамической гипотезы, применительно к задаче распознавания третичной структуры белков, как правило, пользуются некой оптимальной функцией энергии. Существует два различных подхода к построению используемых функций энергии. Первый основывается на физической функции энергии, которая, в принципе, может быть получена в результате рассмотрения различных физических сил взаимодействия между частицами. Второй состоит в построении функции энергии системы «белок — растворитель» исходя из уже имеющихся данных о структуре ранее исследованных белков (часто исследуют статистику парных контактов аминокислот белка, а также структуру его внешней поверхности, контактирующую с раствором). Первый тип функций назовем физически оптимальными функциями энергии (ФОФЭ), второй — статистически оптимальными функциями энергии (СОФЭ).

ФОФЭ основаны на реальных процессах, происходящих в белках, и теоретически способны учитывать все возможные эффекты, важные для предсказания их третичной структуры. Однако при этом ФОФЭ, как правило, слишком громоздки для вычислений, а реальный профиль функции свободной энергии имеет множество локальных минимумов в области, близкой к естественному состоянию белка в растворе, что еще больше затрудняет его вычислительный поиск. Тем не менее периодически появляются новые вычислительные методы сглаживания этой функции в области минимума, решающие данную проблему. Используемые на практике ФОФЭ учитывают молекулярную механику белка и его взаимодействие с раствором. Эти функции являются эмпирическими и приближенными. Исходные данные для их построения получены в результате исследования взаимодействия с растворителем систем, более простых, чем белки, и последующей параметризации. Как правило, ФОФЭ учитывают взаимодействие Ван-дер-Ваальса, энергию взаимодействия с ядром белка, а также содержат ряд компенсирующих слагаемых, например для учета энергии водородных связей и пр.

СОФЭ основываются на статистике, полученной из уже известных белковых структур. Чаще всего используются частотные распределения пар контактирующих белковых остатков, а с увеличением количества данных стало возможным использование частотных распределений пар контактирующих атомов. Считается, что задания частот пар достаточно для построения эффективных моделей предсказания структуры белков. В некоторых случаях к этим данным добавляют и другие составляющие (частоты распределения контактных троек и четверок, вероятности двугранных углов главной и побочной цепей), полезные при вычислении СОФЭ. Такая свобода действий в создании СОФЭ является одновременно и слабой и сильной стороной метода. До настоящего времени подход СОФЭ, широко применяемый на практике, не обоснован теоретически и не имеет какой-либо методологии, общей структуры и классификации. В целом к методам СОФЭ можно отнести любой метод построения функции энергии белка, в котором используется статистическая информация о других белках.

2.1. Статистические функции энергии белка. Суть определения СОФЭ состоит в построении распределений вероятностей пространственного расположения частей белка в определенных конфигурациях. Например, отношение типа «снаружи-внутри» используется для статистического моделирования свойства гидрофобности. Можно также строить СОФЭ на любых фиксированных геометрических конфигурациях, собирая статистику о типах частиц и их расположении в этих конфигурациях. На основе вероятности нахождения частиц в определенных конфигурациях можно построить функцию потенциальной энергии, воспользовавшись уравнением Больцмана

$$\Delta G = -RT \ln (p_{\text{obs}} / p_{\text{exp}}),$$

где p_{obs} — наблюдаемая вероятность определенной конфигурации, p_{exp} — ожидаемая вероятность наблюдения этой конфигурации [2, 3].

Таким образом, возможно использование целого ряда различных пространственных конфигураций одновременно и последующее их суммирование с эмпирически подобранными весовыми коэффициентами. Этот подход чаще всего применяется параллельно с методом Монте-Карло.

Преимуществом СОФЭ над ФОФЭ является их меньшая чувствительность к незначительным смещениям частиц в пространственной модели белка. Еще одно неоспоримое достоинство СОФЭ — их статистическая основа, позволяющая учитывать любые физические явления и эффекты, включая и не известные в настоящее время.

К недостаткам СОФЭ следует отнести частое возникновение шумов, вызванных неточностями конкретных методов. Например, в модели СОФЭ, использующей частоты пар контактирующих аминокислот, будут возникать погрешности, если между этими аминокислотами находится атом металла ядра. Однако если причину шумов удается найти, то ее, как правило, можно легко устранить без существенных изменений самой модели СОФЭ. Таким образом, проблема сводится к выявлению всех источников шумов.

Отдельно следует отметить, что при построении СОФЭ не учитывают все частицы белка, а исследуют взаимодействие лишь отдельных их групп, наиболее сильно влияющих на структуру белка. Выбор этих групп также определяет эффективность метода.

2.2. Физические функции энергии белка. Каждая группа исследователей разрабатывает свой конкретный вид ФОФЭ. При этом члены, входящие в формулу ФОФЭ, могут существенно отличаться у каждой научной группы. Опишем обязательные составляющие части ФОФЭ. Члены в функции энергии можно разделить на две группы: энергия связи и остальные. К энергии связи обычно относят взаимодействие двух или четырех атомов, соединенных ковалентной связью; их роль заключается в ограничении допустимой области длин и углов связей вблизи положения равновесия. К остальным членам относят потенциал Леннарда–Джонса (в виде взаимодействия Ван-дер-Ваальса между отдельными атомами) и закон Кулона. Параметры для связанных и не связанных членов получают с помощью квантовых вычислений на основе термодинамических, кристаллографических и спектроскопических экспериментов на реальных молекулах.

Рассмотрим отдельные члены, входящие в ФОФЭ, уделяя внимание вопросам, возникающим при попытке их точного вычисления [3].

Энергия связей. Энергии ковалентных связей групп атомов варьируются в широком диапазоне в зависимости от угла и расстояния. Таким образом, функция энергии связи не несет смысловой нагрузки, пока структура молекулы не ограничена пространственными углами и не задано хотя бы примерное расположение частиц.

Потенциал Леннарда–Джонса. Взаимодействие Ван-дер-Ваальса моделируется с помощью потенциала Леннарда–Джонса:

$$V(r) = 4\varepsilon \left[\left(\frac{\sigma}{r} \right)^{12} - \left(\frac{\sigma}{r} \right)^{6} \right].$$

Первый член в скобках уравнения очень чувствителен к небольшим смещениям частиц. Дискретное моделирование делает невозможным точное нахождение минимума этого потенциала.

Во время моделирования нередко возникают перекрытия и частичные совмещения двух частиц, что приводит к сильно завышенным значениям потенциала. Эта проблема решается либо линеаризацией функции энергии для малых расстояний между частицами, либо понижением степени в уравнении и подбором оптимальных коэффициентов в формуле. Современные возможности вычислительной техники не позволяют использовать точную формулу расчета потенциала Леннарда–Джонса.

Сольватация. Энергия сольватации состоит из двух слагаемых: электростатического взаимодействия между атомами белка и растворителем, гидрофобного эффекта взаимодействия контактирующих атомов белка и растворителя (поверхностная энергия).

Точный расчет энергии сольватации подразумевает расчет взаимодействия атомов белка с множеством единичных атомов растворителя. Эта задача имеет чрезмерную вычислительную сложность, и на практике используются упрощенные модели. Как пример, растворитель заменяется однородной средой с определенным однородным коэффициентом поляризации и поверхностным натяжением. Все расчеты, таким образом, основаны для поляризационного эффекта на законе Кулона и потенциале Леннарда–Джонса, а для гидрофобного — на площади поверхности контакта белка с растворителем. Важным здесь является применимость принципа суперпозиции для этих законов, позволяющая просто суммировать различные взаимодействия для каждого отдельного атома белка.

Модель для расчета сольватации, приведенная в качестве примера, является чрезмерно упрощенной и в настоящее время признана непригодной для точного предсказания пространственной структуры белков.

Гидрофобный эффект. При формировании третичной структуры белка важную роль играет полярность аминокислотных остатков. Неполярные аминокислоты формируют ядро белковой глобулы, избегающее контактов с молекулами воды. Полярные остатки формируют оболочку, контактирующую с молекулами воды, окружающими глобулу. Поместив аминокислотные остатки первичной структуры белка в узлах трехмерной решетки с длиной ребра, равной длине пептидной связи, можно оценить количество контактов между неполярными остатками (HH-контакты). Задача сводится к нахождению такой конфигурации ω, которая бы максимизировала количество HH-контактов. Энергия конфигурации ω записывается в виде

$$E = -n_{\mathrm{HH}}(\omega),\ \omega \in \Omega,$$

где $n_{\mathrm{HH}}(\omega)$ — количество HH-контактов в конфигурации ω, Ω — множество всех возможных конфигураций белка с заданной первичной структурой. Для минимизации энергии используется, как правило, весь арсенал современных методов оптимизации.

Гидрофобный эффект обычно моделируется как некая функция, пропорциональная площади поверхности контакта белка с растворителем. Тем не менее существует ряд эффектов, замеченных в реальных белках, для которых такая модель не работает. Например, в случае, когда различные слои белка разделены лишь мономолекулярным слоем растворителя, атомы белка по обе части растворителя взаимодействуют между собой.

Существующие модели учета подобных эффектов неточны, ресурсоемки и, более того, требуют дополнительной аппроксимации и оптимизации алгоритма.

Вода. Возможны ситуации, когда отдельные молекулы воды плотно связываются с белком. В таком случае стандартная модель «белок — растворитель» неприменима. Необходимо разрабатывать специальные алгоритмы предсказания такого рода связывания и его учета в ФОФЭ.

Водородные связи. В классическом случае водородные связи моделируются как электростатическое взаимодействие между двумя диполями (например, C=O и N–H). При этом все четыре атома в модели обычно помещаются на одной прямой. В работе [1] показано, что в некоторых случаях это допущение неверно, и разработана модель расчета для «искривленной» связи, а также статистический метод предсказания появления подобного рода «искривлений».

Поляризация растворителя и квантовые эффекты. В основе подавляющего большинства ФОФЭ лежат модели, рассматривающие фиксированные заряды атомов. Тем не менее при взаимодействии с растворителем на атомах вследствие квантовых эффектов возникают наведенные диполи, что в конечном счете влияет на качество предсказания структуры белка.

Построение точной модели с учетом квантовых эффектов невозможно из-за огромной вычислительной сложности. В качестве компромисса возможно использование квантовых расчетов для уточнения наиболее важных участков структуры белка.

Таким образом, каждый отдельный член функции энергии требует от исследователей решения целого ряда задач и принятия специфических решений, что в итоге приводит к ситуации, когда каждая группа разработчиков использует определенную функцию энергии.

Отдельно следует более подробно обсудить, по мнению большинства исследователей, важный вопрос вычисления энергии сольватации. Предполагается, что, основываясь только на внутреннем молекулярном взаимодействии между атомами белка, невозможно построить процедуру распознавания его структуры. Для более точного предсказания структуры необходимо учитывать взаимодействие белка с растворителем. При этом гидрофобный эффект, вследствие которого отдельные аминокислоты белка стремятся занять положение ближе к его ядру, играет второстепенную роль. Основной вклад в функцию энергии дает не гидрофобный эффект, а энергия, необходимая для десольватации белка.

Вычисление энергии сольватации белка является наиболее трудным в процессе применения ФОФЭ. Основным уравнением, используемым для нахождения свободной энергии сольватации, является уравнение Пуассона–Больцмана (ПБ). К сожалению, это уравнение из-за его сложности невозможно использовать в упрощенной форме в машинных вычислениях.

Таким образом, большинство применяемых ФОФЭ состоит их трех слагаемых: межмолекулярного взаимодействия, ПБ и члена, учитывающего гидрофобный эффект. При этом не рассматриваются колебательные процессы в молекуле белка и любая другая кинетика.

Ввиду большой вычислительной сложности уравнения ПБ разработаны другие упрощенные подходы. Наиболее перспективен подход, основанный на обобщенной модели Бора (ОМБ), в которой для вычисления боровского радиуса вместо уравнения ПБ применяется упрощенная модель электростатического взаимодействия атомов. Большинство эффективных ФОФЭ, основанных на ОМБ, используют функции CHARMM — наиболее известной программы молекулярного моделирования, применяемой для систем с большим количеством частиц и ставшей своего рода стандартом.

В качестве примера подхода, не использующего ПБ или ОМБ, можно привести модель гауссовского исключения сольватации. Если этот метод совмещают с применением функции энергии CHARMM, то его называют EEF1. В настоящее время EEF1, а также методы, использующие ОМБ, полностью реализованы в CHARMM.

2.3. Применение биофизических фильтров и карт Рамачандрана. В методах минимизации энергии по известной первичной структуре белка генерируют третичную структуру и проверяют значение ее энергии. Затем изменяют в этой третичной структуре какие-то параметры и подсчитывают энергию у нового варианта. Если его энергия меньше, чем у предыдущего, то работают с новым вариантом структуры, в противном случае возвращаются к предыдущему варианту. Затем в текущем варианте структуры снова проводят изменения и т.д.

Поскольку вычисление энергии весьма затруднительно, то еще до этапа ее подсчета из рассмотрения исключают варианты структуры, не имеющие физического смысла. Такую проверку делают с помощью карт Рамачандрана и так называемых биофизических фильтров. Карты Рамачандрана показывают разрешенные и запрещенные конформации (разрешенные и запрещенные двугранные углы химических связей) для аминокислотных остатков белка.

Двугранный (торсионный) угол. Для описания взаимного расположения атомов линейной четырехатомной индивидуальной молекулы (или входящего в состав более сложной молекулы линейного четырехатомного фрагмента) A–B–C–D используется такой геометрический параметр, как двугранный (называемый также торсионным) угол. Это угол между плоскостью, в которой лежит фрагмент A–B–C, и плоскостью, в которой лежит фрагмент B–C–D, т.е. он отражает характер взаимного расположения этих трехатомных фрагментов. На рис. 1 двугранный угол обозначен ψ, величины валентных углов трехатомных фрагментов A–B–C и B–C–D — соответственно γ_1 и γ_2.

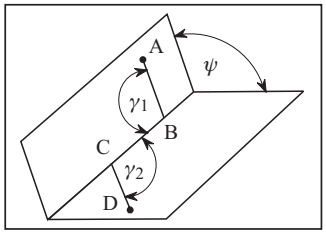

Рис. 1. Двугранный (торсионный) угол

Наиболее наглядно двугранный угол может быть изображен при рассмотрении A–B–C–D вдоль направления связи B–C. Тогда он будет виден как угол между связями B–A и C–D.

Торсионные углы связей полипептидной цепочки белка. В молекуле любого белка можно выделить линейную многоатомную цепочку, проходящую через все аминокислоты белка. В эту цепочку из каждого аминокислотного остатка входят по три атома (N_i, C_i^α, C_i, где i — номер аминокислотного остатка в аминокислотной последовательности белка). Связи между этими атомами обозначаются следующим образом: $N_i - C_i^\alpha$, $C_i^\alpha - C_i$, $C_i - N_{i+1}$, $N_{i+1} - C_{i+1}^\alpha$, $C_{i+1}^\alpha - C_{i+1}$, $C_{i+1} - N_{i+2}$ и т.д.

Торсионный угол, описывающий вращение вокруг связи $N-C^\alpha$, обозначают ϕ, вокруг связи $C^\alpha - C$ — как ψ, а описывающий вращение вокруг пептидной связи C–N — как ω. Символы ϕ_i, ψ_i, ω_i используются для обозначения торсионных углов в пределах i-го аминокислотного остатка в случаях ϕ и ψ и между i-м и $(i+1)$-м остатком в случае ω. Торсионный угол ϕ_i определяется последовательностью атомов C_{i-1}, N_i, C_i^α, C_i, угол ψ_i — последовательностью N_i, C_i^α, C_i, N_{i+1}, а угол ω_i — последовательностью C_i^α, C_i, N_{i+1}, C_{i+1}^α (рис. 2). Так как пептидная связь $C_i - N_{i+1}$ частично двойная, угол ω обычно может принимать значения только в окрестностях 0° или 180°.

На рис. 2 показан участок полипептидной цепи с двумя пептидными связями. Границы между аминокислотными остатками обозначены волнистыми линиями. (Цепь показана в наиболее растянутой конформации: $\phi_i = \psi_i = \omega_i = 180°$).

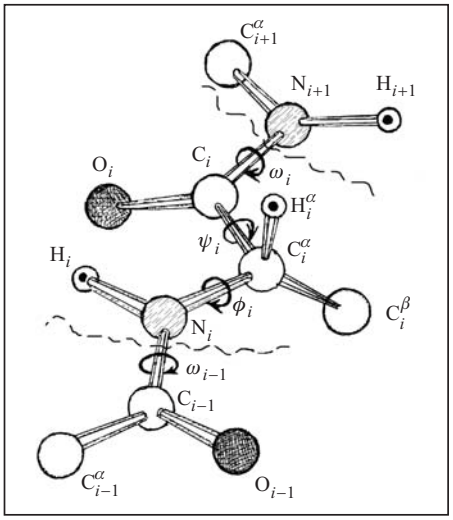

Рис. 2. Торсионные углы между связями полипептидной (белковой) цепи

График Рамачандрана (называемый также картой или диаграммой Рамачандрана, конформационной картой, ϕ/ψ-картой). Этот график показывает выведенные из квантово-химических расчетов разре-

шенные и запрещенные значения двугранных углов ϕ и ψ аминокислот или полученные экспериментально (с помощью рентгеноструктурного анализа существующих белков) частоты встречаемости значений этих углов. На оси абсцисс показываются углы ϕ, на оси ординат — ψ.

Для торсионных углов того или иного вида вторичной структуры белка также могут быть построены карты Рамачандрана.

В методах минимизации энергии на этапе генерации вариантов третичной структуры торсионные углы каждой аминокислоты генерируются с учетом их разрешенных значений на карте Рамачандрана. Конформации, имеющие запрещенные торсионные углы, не генерируются.

Биофизические фильтры. После генерации каждого варианта третичной структуры он может быть проверен биофизическими фильтрами, с помощью которых у проверяемого варианта можно обнаружить несвойственные существующим в природе нативным структурам белков особенности. Если такие особенности находятся, то данный вариант структуры отбрасывается и подсчет его энергии не производится.

Предложены различные биофизические фильтры: по длине протяженности, по радиусу инерции, по пропорции гидрофобности (Hydrophobicity ratio filter), по упакованной фракции (Packing fraction filter) и другие, в которых используются различные биофизические характеристики. По сравнению с другими биофизическими фильтрами наиболее эффективны в выявлении структур, не похожих на природные, фильтры по длине протяженности и по радиусу инерции [4]. Они кратко описаны ниже.

Длина протяженности (Persistence length) — максимальная длина по прямой линии между участками непрерывной полипептидной цепи. Для глобулярного белка ее значения варьируют от 15 Å до 60 Å со средним значением около 40 Å [4]. Этот интервал используется как порог, чтобы выделить конформации, не схожие с природными.

Радиус инерции (Radius of gyration) белка определяется как среднее квадратичное расстояний между каждым атомом белка и их общим центром масс. Радиус инерции глобулярного белка пропорционален $N^{3/5}$ (где N — количество аминокислот) и удовлетворяет следующему равенству:

$$R_g = \alpha \times N^{3/5} + \beta.$$

Для того чтобы отличить схожие с нативными структуры от несхожих, обычно используют фиксированное значение $\alpha = 0{,}359$ и значения β в промежутке от 4,257 до 11,257 [4, 5].

2.4. Карты контактов. Контактная карта (или карта контактов) белка — упрощенное представление трехмерной структуры белка. Оно несет информацию только о расстояниях между аминокислотами в трехмерном пространстве и представляет собой булеву симметричную квадратную матрицу M размерности $N \times N$, в которой элемент $M(i, j) = 1$, если расстояние между i-м и j-м аминокислотными остатками белка меньше некоторого порогового значения, и $M(i, j) = 0$ в противном случае, где $i = 1, \ldots, N$ и $j = 1, \ldots, N$ — порядковые номера аминокислот в первичной структуре белка, N — количество аминокислот в белке.

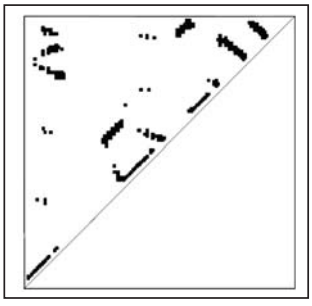

Рис. 3. Пример контактной карты белка

Для построения контактных карт могут использоваться различные определения и пороговые значения расстояния между аминокислотными остатками: расстояние между атомами C^{α} с порогом 6–12 Å, расстояние между атомами C^{β} с порогом 6–12 Å (в этом случае для глицина используется атом C^{α}) или наименьшее расстояние между любыми атомами, но с меньшим порогом 4,5–6 Å.

На рис. 3 приведен пример контактной карты. Контакты (элементы матрицы, равные единице) отмечены темными квадратами. Для удобства показаны элементы только одной из двух полностью симметричных половин матрицы, и только те, для которых $|i-j| \geq 3$.

Взаимное расположение контактов на контактной карте подчиняется следующим закономерностям:

— контакты между аминокислотами одной α-спирали расположены на расстоянии не более нескольких позиций от главной диагонали карты;

— контакты между аминокислотами двух соседних тяжей антипараллельной β-структуры занимают область толщиной в несколько позиций, вытянутую перпендикулярно главной диагонали карты;

— контакты между аминокислотами двух соседних тяжей параллельной β-структуры занимают область толщиной в несколько позиций, вытянутую параллельно главной диагонали карты.

Контактные карты могут использоваться как промежуточное звено для предсказания координат атомов белка в трехмерном пространстве. Этот подход перспективен, так как контакты предсказываются методами машинного обучения легче, чем координаты в трехмерном пространстве, а последние при определенных условиях можно реконструировать по известной контактной карте [6].

Разработано много различных методов для предсказания контактной карты по известной первичной структуре белка. Их можно разделить на две взаимонеисключающие категории:

1) статистические подходы, использующие метод коррелирующих мутаций;

2) подходы, использующие методы машинного обучения.

Первые для предсказания контактов используют информацию о коррелирующих мутациях разных аминокислотных остатков, вторые — методы машинного обучения, такие как нейронные сети, скрытые марковские модели и метод опорных векторов [7]. Для предсказания контактных карт могут применяться также комбинации нескольких методов.

Точность предсказания контактных карт в зависимости от используемого для предсказания подхода может быть связана с качеством множественного выравнивания последовательностей и точностью предсказания вторичных структур. Кроме того, она всегда связана с пропорцией β-листов в белке — контакты в β-листах в β-белках, $\alpha+\beta$-белках и α/β-белках предсказываются с большей точностью, чем контакты между α-спиралью и β-листом или между α-спиралями [8].

Контакты с участием аминокислоты, не входящей во вторичные структуры, между двумя разными α-спиралями, между α-спиралью и β-структурой, и даже между двумя разными β-структурами, не параллельными и не антипараллельными одна другой, как правило, предсказываются значительно хуже, чем контакты между аминокислотами одной вторичной структуры. Интересно, что большинство неверно предсказанных контактов расположены вблизи от настоящих контактов.

Для выявления неправильно предсказанных контактов могут быть использованы правила, основанные на геометрических закономерностях строения α-спиралей и β-структур.

Наибольшая точность предсказания контактов по известной аминокислотной последовательности составила 32%; полученная оценка превосходит точность предсказания координат в трехмерном пространстве по известной аминокислотной последовательности.

Общая задача вычисления координат атомов белка в трехмерном пространстве, совместимых с данной контактной картой, известная как задача о существовании графа дисков единичного радиуса, является *NP*-трудной задачей. В настоящее время разработано несколько эмпирических методов для предсказания координат атомов белка по известной карте контактов. Для построенных по эмпирическим данным картам контактов координаты атомов белка могут быть восстановлены со

средним квадратичным отклонением 1–2 Å, однако по предсказанным картам контактов они прогнозируются с высоким отклонением более 3 Å [6].

3. ВЕРОЯТНОСТНЫЕ МОДЕЛИ НА ПОСЛЕДОВАТЕЛЬНОСТЯХ

При решении задачи предсказания структуры белка применяются вероятностные модели на последовательностях. Модели строятся по информации из обучающих выборок, в качестве которых используются открытые банки данных белковых структур. Для заданной последовательности аминокислот, или наблюдений, требуется найти наиболее вероятную последовательность состояний. Состояниями могут быть типы вторичной структуры или торсионные углы в зависимости от конкретной задачи.

Обозначим последовательность наблюдений $x = (x_1, \ldots, x_n)$, $x_i \in A_X$, а последовательность искомых состояний $y = (y_1, \ldots, y_n)$, $y_i \in A_Y$. Здесь A_X — конечное множество значений одного наблюдения, A_Y — конечное множество значений одного состояния. Кроме того, обозначим $X = A_X^n$ множество всех возможных значений последовательности наблюдений длины n, $Y = A_Y^n$ — множество всех возможных значений последовательности состояний длины n.

Предполагается, что существует совместное распределение вероятности на множестве наблюдений и состояний $P(x, y)$, $x \in X$, $y \in Y$; обозначим его $P(v)$, $v \in V$, $V = X \times Y$. Задача ставится следующим образом: с помощью обучающей выборки необходимо найти $\arg\max_{y \in Y} P(y \mid x)$, $x \in X$. Для решения этой задачи используются модели Маркова со скрытыми параметрами (HMM — Hidden Markov Models) и условные случайные поля (CRF — Conditional Random Fields). Эти подходы основаны на представлении многомерного совместного распределения $P(v)$, $v \in V$, в виде произведения более простых распределений — факторов, характеристики которых, в отличие от исходного распределения, можно установить из имеющихся обучающих выборок.

Модели Маркова со скрытыми параметрами (ММСП). Предполагается, что последовательность состояний $y = (y_1, \ldots, y_n)$ описывается цепью Маркова, например, первого порядка. Задано начальное распределение вероятности состояний $P(y_i)$ и вероятности переходов между состояниями $P(y_i \mid y_{i-1})$, $y_i, y_{i-1} \in A_Y$. Кроме того, предполагается, что наблюдения x_i независимы. Заметим, что с помощью критерия χ^2 гипотеза о независимости аминокислотной последовательности x_i легко отвергается [9]. Таким образом, вопрос относительно адекватности описываемой ниже модели остается за скобками. Вероятность значения x_i зависит только от значения текущего состояния y_i с заданным распределением $P(x_i \mid y_i)$, $y_i \in A_Y$, $x_i \in A_X$. Это позволяет записать совместное распределение вероятности в виде

$$P(v) = p(y_1)p(x_1 \mid y_1) \prod_{i=2}^{n} p(y_i \mid y_{i-1}) p(x_i \mid y_i).$$

На основе совместного распределения с помощью ММСП решаются следующие задачи.

- Задана последовательность наблюдений $x = (x_1, \ldots, x_n)$ и модель λ. Необходимо найти наиболее вероятную, в определенном смысле, последовательность состояний $y = (y_1, \ldots, y_n)$.
- Дана обучающая выборка. Необходимо подобрать модель λ, которая бы максимизировала совместное правдоподобие последовательностей наблюдений и состояний.

Эти задачи сводятся к нахождению оптимальных параметров модели λ по обучающей выборке и определению с помощью полученной модели наиболее вероятной структуры, которую имеет заданная аминокислотная последовательность белка.

Существует несколько подходов к решению первой задачи, которые отличаются определением «наиболее вероятной» последовательности состояний $y=(y_1,\dots,y_n)$, соответствующих последовательности наблюдений $x=(x_1,\dots,x_n)$. Например, можно выбирать последовательность y таким образом, чтобы каждое состояние y_i имело наибольшую вероятность при заданном значении наблюдения x_i. Обычно используют другой критерий, который требует нахождения наиболее вероятной последовательности y при заданной всей последовательности x, т.е.

$$\arg\max_{y\in Y} P_\lambda(y|x),\ \ y\in Y.$$

Эта задача эквивалентна максимизации $\arg\max_{y\in Y} P_\lambda(x,y)$, учитывая, что последовательность x задана. Последняя задача решается с помощью алгоритма Витерби [10], который заключается в следующем.

Формула

$$\delta_t(y')=\max_{y_1,\,y_2,\dots,\,y_{t-1}} P(y_1,\dots,y_{t-1},y_t=y',x_1,\dots,x_t),\ \ y'\in A_y,$$

обозначает максимальное значение вероятности последовательности первых t состояний и наблюдений при заданном значении состояния $y_t=y'$ в момент времени t и заданных значениях наблюдений $x_1,\dots,x_t$. Значения δ_i, $i=2,\dots,n$, находятся рекурсивно,

$$\delta_{t+1}(y')=[\max_{y_t\in A_y}\delta_t(y_t)P(y_{t+1}=y'\,|\,y_t)]P(x_{t+1}\,|\,y_{t+1}=y'),$$

что позволяет найти последовательность состояний $y=(y_1,\dots,y_n)$.

Вторая каноническая задача ММСП — нахождение оптимальной модели по обучающей выборке — решается путем максимизации совместного правдоподобия, что эквивалентно решению задачи глобальной оптимизации. Оптимальной процедуры нахождения параметров модели λ в настоящее время не существует. В качестве приближенных методов обычно используют метод Баума–Уэлша [11] или градиентные методы [12].

Графическое представление. Прежде чем перейти к описанию методов условных случайных полей (УСП), рассмотрим графические модели, которые применяются для описания многомерных распределений. Суть графического подхода заключается в представлении совместного многомерного распределения в виде графов: графа зависимостей и графа факторов.

Рассмотрим граф зависимостей $G_d=(V_d,E_d)$, где V_d — множество вершин, каждая из которых соответствует случайной величине v_i, входящей в $v\in V$; E_d — множество ребер графа. Факт отсутствия ребра (v_i,v_j) в множестве E_d выражает условную независимость случайных величин v_i,v_j. Случайные величины a и b называются условно независимыми при заданном значении третьей случайной величины c, если справедливо равенство $P(a\,|\,b,c)=P(a\,|\,c)$.

Граф зависимостей задает внутреннюю структуру совместного распределения $P(v)$, $v\in V$, и не несет информации о количественных характеристиках взаимосвязей между случайными величинами v_i. Ребра в графе зависимостей G_d могут быть направленными или ненаправленными. В зависимости от этого различают два разных типа графических моделей: ММСП относятся к направленным моделям, а УСП — к ненаправленным.

Для количественного описания зависимостей между случайными величинами v_i используется так называемый граф факторов $G_f=(V_f,E_f)$; V_f содержит помимо вершин, соответствующих случайным величинам v_i, еще вершины, соответству-

ющие факторам $\psi_k(v_k, N(v_k))$ (рис. 4). В направленных моделях фактор ψ_k является условным распределением вероятности значений v_k при заданных значениях множества родительских вершин $N(v_k)$. Вершина v_i принадлежит множеству родительских вершин v_k, если множество ребер E_d содержит направленное ребро (v_i, v_k), т.е.

$$v_i \in N(v_k) \Leftrightarrow \exists\,(v_i, v_k) \in E_d.$$

Исходное распределение $P(v)$ записывается в виде произведения факторов

$$P(v) = \prod_{k=1}^{K} \psi_k(v_k, N(v_k)),\ v \in V,$$

где K — количество всех факторов в модели.

Рассмотрим для примера ММСП на последовательности из трех состояний и трех наблюдений $v = (x_1, x_2, x_3, y_1, y_2, y_3)$; соответствующие графы G_d и G_f изображены на рис. 4, *а,б*.

Совместное распределение вероятности v записывается в виде

$$P(v) = P(y_1)P(x_1 \mid y_1)P(y_2 \mid y_1)P(x_2 \mid y_2)P(y_3 \mid y_2)P(x_3 \mid y_3),$$

или то же самое в терминах факторов

$$P(v) = \psi_1(y_1)\psi_4(x_1, y_1)\psi_2(y_2, y_1)\psi_5(x_2, y_2)\psi_3(y_3, y_2)\psi_6(x_3, y_3).$$

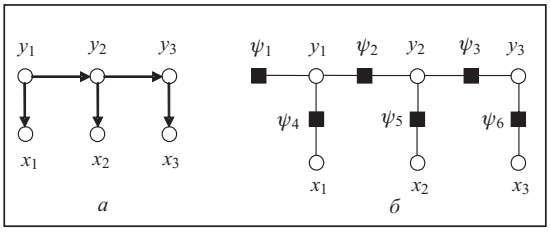

Рис. 4. Графы G_d и G_f

Условные случайные поля (УСП). Модели УСП относятся к ненаправленным моделям, что влечет некоторые изменения при подсчете совместной вероятности. В ненаправленных моделях понятие родительской вершины теряет свой смысл, поэтому факторы $\psi_c(V^c)$ определяются на максимальных кликах $c \in C$ графа зависимостей G_d, V^c — множество вершин, входящих в клику $c \in C$ [13]. Условное распределение задается формулой

$$P(y \mid x) = \frac{1}{Z(x)} \prod_{c \in C} \psi_c(V^c).$$

Факторы ψ_c имеют вид

$$\psi_c(x, y) = \exp\left(\sum_{i=1}^{m} \lambda_i f_i(V^c) \right),$$

где $f_i(V^c)$, $i = 1, \ldots, m$, — произвольный набор неотрицательных интегрируемых по Лебегу функций, которые не обязательно должны иметь вероятностную интерпретацию. Отсутствие направлений дает возможность использовать широкий набор функций в качестве факторов и не ограничиваться только условными вероятностями, как в случае с ММСП. Для того чтобы в конечном итоге получить вероятностную меру $P(y \mid x)$, необходимо проводить нормализацию. Нормализующий множитель $Z(x)$ имеет вид

$$Z(x) = \sum_{y \in Y} \prod_{c \in C} \psi_c(x^c, y^c).$$

Нахождение $Z(x)$ — сложная вычислительная процедура, которая сводится к вычислению многомерного интеграла. Это своего рода плата за свободный выбор функций f_i, определяющих факторы ψ_c. В направленных моделях такой проблемы не возникает, так как в качестве факторов используются условные вероятности и нормализации не требуется.

Модели максимальной энтропии. Общий вид факторов ψ_c УСП выводится из принципа максимальной энтропии. Согласно этому принципу при наличии неполной информации о распределении следует выбирать распределение, наиболее равномерное относительно имеющейся информации, т.е. максимизирующее энтропию. Иными словами, любое другое распределение (с меньшей энтропией) несет в себе информацию, которой не было в исходных данных о распределении.

В случае условного распределения $P(y|x)$ используется условная энтропия $H[P(y|x)] = -\sum_{(x,\, y)\in V} P(x, y)\log P(y|x)$.

Необходимо найти распределение, которое максимизирует энтропию и «согласуется» с обучающей выборкой T:

$$P^*(y|x) = \arg\max_{P(y|x)\in\Pi} H[P(y|x)].$$

Здесь П — множество всех моделей. Для удобства будем полагать, что x, y — значения одного наблюдения и одного состояния.

«Согласованность» искомого распределения с обучающей выборкой представляется с помощью набора неотрицательных интегрируемых по Лебегу функций $f_i(x, y)$, $i = 1,\ldots, m$.

Первые m ограничений на распределение $P(y|x)$ заключаются в том, что эмпирические ожидания $\hat{E}f_i(x, y)$ должны совпадать с прогнозируемыми ожиданиями $Ef_i(x, y)$:

$$\hat{E}f_i(x, y) = Ef_i(x, y),\ 1 \le i \le m.$$

Эмпирическое ожидание записывается в виде

$$\hat{E}f_i(x, y) = \frac{1}{N}\sum_{(x,\, y)\in T} f_i(x, y),$$

где T — множество обучающих примеров мощности $|T| = N$. Прогнозируемое ожидание записывается в виде

$$Ef_i(x, y) = \sum_{(x,\, y)\in V} P(x)P(y|x)f_i(x, y).$$

Для того чтобы эффективно вычислить $Ef_i(x, y)$, вместо $P(x)$ используется эмпирическое распределение $\hat{P}(x)$. Тогда

$$Ef_i(x, y) \approx \frac{1}{N}\sum_{x\in T}\sum_{y\in Y} P(y|x)f_i(x, y). \qquad (1)$$

В приложениях множество возможных состояний Y обычно не так велико по сравнению с X, что позволяет эффективно провести суммирование в (1). Например, в задаче распознавания вторичной структуры белка количество состояний — три (α-спираль, β-слой, *coil*), тогда как количество возможных значений наблюдений — 20.

Еще одно ограничение на искомую модель $P(y|x)$ заключается в том, чтобы $P(y|x)$ являлась вероятностной мерой, т.е. $P(y|x) \ge 0\ \ \forall x \in X,\ \ y \in Y$ и $\sum_{y\in Y} P(y|x) = 1\ \ \forall x \in X$.

Нахождение $P^*(y|x)$ при таких ограничениях является задачей оптимизации с ограничениями. Функция Лагранжа имеет вид

$$\Lambda(P, \lambda) = H[P(y|x)] + \sum_{i=1}^{m} \lambda_i (E(f_i) - \hat{E}(f_i)) + \lambda_{m+1}\left(\sum_{y\in Y} P(y|x) - 1\right),$$

где $\lambda = (\lambda_1, \ldots, \lambda_n)$. Используя эмпирическое распределение $\hat{P}(x)$ при вычислении $H[P(y|x)]$, получают

$$P_\lambda^*(y|x) = \frac{1}{Z_\lambda(x)} \exp\left(\sum_{i=1}^{m} \lambda_i f_i(x, y)\right),$$

$$Z_\lambda(x) = \sum_{y \in Y} \exp\left(\sum_{i=1}^{m} \lambda_i f_i(x, y)\right). \qquad (2)$$

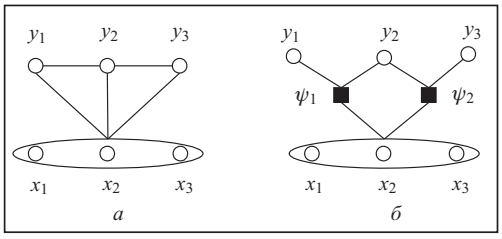

Рис. 5. Графы зависимостей и факторов

УСП на последовательностях. Рассмотрим УСП на конкретном примере. Граф зависимостей, изображенный на рис. 5, *а*, состоит из трех вершин, имеющих состояния y_1, y_2, y_3, и одной вершины, соответствующей всей последовательности наблюдений (x_1, x_2, x_3). Любая пара соседних состояний y_i, y_{i+1} вместе с наблюдениями (x_1, x_2, x_3) образует клику, на каждой из которых определен фактор $\psi_i(x, y)$. Граф факторов изображен на рис. 5, *б*.

Условная вероятность $P(y|x)$ для последовательностей $y \in Y$, $x \in X$ длины n записывается в виде

$$P(y|x) = \frac{1}{Z(x)} \prod_{j=1}^{n-1} \psi_j(x, y), \quad x \in X, \; y \in Y,$$

с нормализующей константой

$$Z(x) = \sum_{y \in Y} \prod_{j=1}^{n-1} \psi_j(x, y)$$

и факторами

$$\psi_j(x, y) = \exp\left(\sum_{i=1}^{m} \lambda_i f_i(y_j, y_{j+1}, x, j)\right).$$

Индекс j в функциях f_i, в отличие от формулы (2), необходим, поскольку здесь x — последовательность случайных величин. Функции f_i на практике определяются не на всех элементах последовательности x, а на некотором подмножестве, которое определяется относительно текущего индекса j. Функции f_i могут иметь вид индикаторных функций некоторых событий на множестве $V = X \times Y$. Например, в контексте задачи распознавания вторичной структуры белка можно представить индикаторную функцию события, состоящего в том, что состояние y_j соответствует α-спирали, состояние y_{j+1} — *coil*, а на месте x_j находится аминокислота метионин

$$f_i(y_j, y_{j+1}, x, j) = \begin{cases} 1, & \text{если } y_j = \alpha, \; y_{j+1} = c, \; x_j = \text{M}, \\ 0 & \text{в противном случае.} \end{cases}$$

Несмотря на то что функция f_i определена на всей последовательности состояний x, используется лишь небольшое подмножество их значений, в данном примере — только x_j.

Окончательный вид модели УСП на последовательности наблюдений $x = (x_1, \ldots, x_n)$ и состояний $y = (y_1, \ldots, y_n)$ длины n записывается следующим образом:

$$P(y|x) = \frac{1}{Z(x)} \prod_{j=1}^{n-1} \exp\left(\sum_{i=1}^{m} \lambda_i f_i(y_j, y_{j+1}, x, j)\right).$$

Для УСП, как и для ММСП, решаются задача нахождения оптимальных параметров $\lambda = (\lambda_1, \ldots, \lambda_n)$ и задача нахождения наиболее вероятной последовательности состояний при заданной последовательности наблюдений.

Задача нахождения оптимальних параметров λ модели $P_\lambda(y|x)$ решается методом максимального правдоподобия, что, как и в случае с ММСП, сводится к решению задачи глобальной оптимизации. Оптимальной процедуры нахождения параметров модели не существует, применяются градиентные методы или модифицированный метод Баума–Уэлша.

Задача построения последовательности состояний $y = (y_1, \ldots, y_n)$ по последовательности наблюдений $x = (x_1, \ldots, x_n)$ решается с помощью модифицированного метода Витерби.

Сходство ММСП и УСП состоит в том, что многомерное совместное распределение представляется в виде произведения независимых факторов. На основе полученного совместного распределения решается задача обучения или нахождения оптимальных параметров модели и задача нахождения последовательности состояний при заданной последовательности наблюдений.

ММСП относится к направленным моделям — факторы имеют вид условных распределений и определяются на вершине v_k и множестве родительских вершин v_k — $N(v_k)$.

УСП относится к ненаправленным моделям — факторы определены на максимальных кликах графа зависимости и в общем случае не имеют вероятностного смысла, что обусловливает необходимость нормализации. Общий вид факторов выводится из принципа максимальной энтропии.

4. ЗАКОНОМЕРНОСТИ ЗАПИСИ ГЕНЕТИЧЕСКОЙ ИНФОРМАЦИИ В ГЕНОМАХ И БЕЛКАХ

Генетическая информация клетки хранится в хромосомах, представляющих собой, согласно известной модели Уотсона–Крика, двойную цепочку ДНК. Каждая цепочка состоит из нуклеотидных звеньев (нуклеотидов, оснований) четырех типов: A, T, C, G. Две цепочки спариваются по закону комплементарности (A соединяется с T, а C — с G) и образуют хромосому. Таким образом, одна цепочка ДНК однозначно определяет цепочку, комплементарную себе, и хромосому в целом.

Несмотря на то что ДНК относительно проста и хорошо изучена химически, структура генома человека чрезвычайно сложна и не все его функции известны. На текущий момент длина законченной геномной последовательности составляет 2851 млн нуклеотидов и содержит 341 пробел общим размером 225 млн оснований. Геном человека включает приблизительно 20–30 тысяч белок-кодирующих генов. В работе [14] приведены сведения о законченных последовательностях и размерах пробелов для каждой хромосомы в геноме человека. Числовые расчеты проводились на последовательностях хромосом, характеристики которых соответствуют данным, указанным в [14].

Соотношения комплементарности в записи оснований по одной нити ДНК исследовались в [15–17], в работе [18] содержится список цитируемой литературы по данному вопросу.

Комплементарность в записи оснований по одной нити ДНК хромосомы означает, что выполняются приближенные соотношения

$$n(\mathrm{A}) \approx n(\mathrm{T}), \;\; n(\mathrm{C}) \approx n(\mathrm{G}), \tag{3}$$

где $n(j)$ — количество оснований j, $j \in \{\mathrm{A}, \mathrm{C}, \mathrm{G}, \mathrm{T}\}$, вычисленных на одной нити.

Заметим, что из комплементарности пар букв по двум нитям ДНК не следует, что количества букв A и T, а также C и G, подсчитанные по одной нити, совпадают между собой. Простой пример: на одной нити содержится 4 млн букв A, 3 млн букв C, 2 млн букв G и 1 млн букв T, тогда на второй нити находится соответственно 4 млн букв T, 3 млн букв G, 2 млн букв C и 1 млн букв A. Таким образом, комплементарность по двум нитям выполняется, а по одной нити нет. Из соотноше-

ний (3) вытекает, что молекулярный вес обеих нитей примерно одинаков. Этот момент является важным для упаковки ДНК, в противном случае из-за возникающих напряжений молекула ДНК могла бы разорваться.

Вычисления показали, что частоты комплементарных оснований A и T, а также C и G, подсчитанные по одной нити ДНК, совпадают на всех хромосомах (геном человека, шимпанзе, мыши, рыбы Tetraodon, растений, бактерий и т.д.) [15].

Для пар оснований выполняются соотношения комплементарности

$$n(ij) \approx n(\bar{j}\bar{i}), \tag{4}$$

где $i, j \in \{\mathrm{A}, \mathrm{C}, \mathrm{G}, \mathrm{T}\}$, $\overline{\mathrm{A}} = \mathrm{T}$, $\overline{\mathrm{C}} = \mathrm{G}$, $\overline{\mathrm{T}} = \mathrm{A}$, $\overline{\mathrm{G}} = \mathrm{C}$. Заметим, что пары AT, TA, CG и GC не представлены в (4), поскольку они сами себе антикомплементарны [15, 16].

Запись и считывание оснований у первой нити хромосомы ДНК выполняется слева направо в направлении $5' \to 3'$, а у комплементарной второй нити — в направлении $5' \to 3'$ справа налево (рис. 6).

Известно, что соотношения

$$\hat{p}(ij) = \frac{n(i,j)}{n(i)}, \tag{5}$$

где $n(ij)$ — число пар (ij), $i, j \in \{\mathrm{A}, \mathrm{C}, \mathrm{G}, \mathrm{T}\}$, $n(i)$ — число оснований i в цепи хромосомы, представляют собой оценки переходных вероятностей для однородных цепей Маркова. В [18] показано, что для длинных цепей оценки (5) сходятся по вероятности к значениям переходных вероятностей.

Из соотношений комплементарности (3), (4) вытекает, что вторая комплементарная нить в направлении $5' \to 3'$ имеет такие же оценки переходных вероятностей $\hat{p}(ij)$, что и исходная первая нить (на рис. 6 представлена пара AC и антикомплементарная ей пара GT). Отсюда следует, что вероятности двух противоположных нитей хромосомы, подсчитанные в модели однородной цепи Маркова на основе оценок переходных вероятностей (5), совпадают.

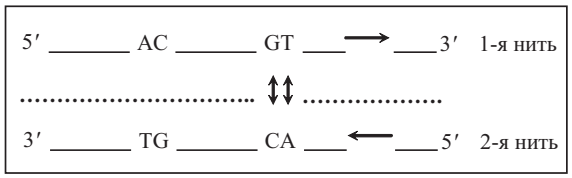

Рис. 6. Условная запись двух нитей хромосомы

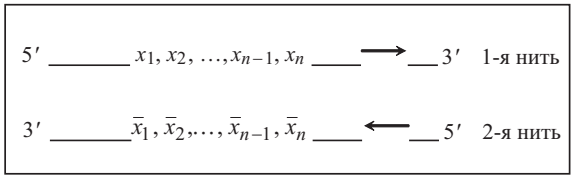

Рис. 7. Комплементарность нуклеотидных последовательностей

Пусть $x_1, x_2, \ldots, x_{n-1}, x_n$ — конечная последовательность оснований, записанных на первой нити, тогда $\bar{x}_n, \bar{x}_{n-1}, \ldots, \bar{x}_2, \bar{x}_1$ — комплементарная ей последовательность оснований, записанных на второй нити (рис. 7).

Для однородной цепи Маркова порядка 1 выполняется следующее важное утверждение.

Лемма. Оценка вероятности последовательности $x_1, x_2, \ldots, x_{n-1}, x_n$ совпадает с оценкой вероятности последовательности $\bar{x}_n, \bar{x}_{n-1}, \ldots, \bar{x}_2, \bar{x}_1$, т.е.

$$\hat{p}(x_1, x_2, \ldots, x_{n-1}, x_n) = \hat{p}(\bar{x}_n, \bar{x}_{n-1}, \ldots, \bar{x}_2, \bar{x}_1). \tag{6}$$

Вероятность однородной цепи Маркова определяется соотношением

$$p(x_1, x_2, \ldots, x_{n-1}, x_n) = p(x_1)p(x_1, x_2) \ldots p(x_{n-1}, x_n), \tag{7}$$

где $p(x_1)$ — вероятность начального состояния, $p(x_{i-1}, x_i)$ — переходные вероятности, $i = 1, 2, \ldots, n$.

Заменив вероятность начального состояния частотой, а переходные вероятности $p(x_{i-1}, x_i)$ в (7) — их оценками (5), получим соотношение (6). Отсюда следует,

что вероятности двух противоположных нитей, подсчитанные для модели однородной цепи Маркова, совпадают.

Кодоны (тройки оснований) связаны соотношениями комплементарности

$$n(i,j,k) \approx n(\bar{k},\bar{j},\bar{i}), \tag{8}$$

где $n(i,j,k)$ — число троек оснований (i,j,k), а $(\bar{k},\bar{i},\bar{j})$ — антикодон кодона (i,j,k). Для 64 триплетов получаем 32 соотношения (4) типа кодон–антикодон. Соотношения комплементарности вида (8) выполняются также для более длинных последовательностей оснований [15, 16].

Оценки переходных вероятностей для цепей Маркова порядка 2 определяются соотношениями

$$\hat{p}(i,j,k) = \frac{n(i,j,k)}{n(i,j)}, \tag{9}$$

где $n(i,j,k)$ — количество троек оснований (i,j,k), а $n(i,j)$ — количество пар (i,j), $i,j,k \in \{\mathrm{A},\mathrm{C},\mathrm{G},\mathrm{T}\}$.

Из соотношений комплементарности (8) заключаем, что оценки переходных вероятностей (9) для обеих нитей, подсчитанные в направлении $5' \to 3'$, совпадают. Легко показать, что результат леммы справедлив и для цепей Маркова порядка 2.

Аминокислотная последовательность белка получается путем трансляции четырехбуквенного алфавита оснований в двадцатибуквенный алфавит аминокислотных остатков. Генетический код образует функцию, которая переводит непересекающиеся тройки оснований в одну из аминокислот. Синтез белков выполняется по двум нитям в направлении $5' \to 3'$. Соотношения комплементарности вида (8) выполняются также для непересекающихся троек оснований и шестерок, состоящих из непересекающихся троек. Поэтому, рассуждая формально, можно сделать вывод, что аминокислотные последовательности белков, синтезированных по первой нити, имеют такие же оценки переходных вероятностей (вида (5)), что и белки, которые синтезируются по второй нити.

Геномы бактерий имеют сравнительно простую структуру: белок-кодирующие участки не прерываются некодирующими вставками — интронами. Эта особенность бактериальных геномов позволяет выделять и отдельно анализировать аминокислотные последовательности белок-кодирующих участков. Численные расчеты, проведенные на геномах бактерий, подтвердили представленный выше вывод. В табл. 1 приведены частоты аминокислот и отдельных пар аминокислот в ДНК бактерии.

Проблема прогнозирования пространственной структуры белков обсуждалась в [19–21], там же приведена постановка задачи предсказания вторичной структуры белков на основе применения эффективных байесовских процедур распознавания на цепях Маркова. Имеется первичная последовательность аминокислот белка, необходимо определить ее вторичную структуру: поставить в соответствие каждой аминокислоте один из двух возможных типов регулярной структуры (α-спираль, β-слой) или ее отсутствие, т.е. нерегулярность (*coil*).

Байесовская процедура распознавания на цепях Маркова. В работах [19–21] исследовалась процедура предсказания вторичной структуры одиночной аминокислоты на основе известной формулы Байеса

$$P(f \mid x_1, x_2, \ldots, x_n) = \frac{P(x_1, x_2, \ldots, x_n \mid f) P(f)}{P(x_1, x_2, \ldots, x_n)}. \tag{10}$$

Здесь f — состояние аминокислоты, число классов f — 60, так как 20 — количество аминокислот, 3 — число вторичных структур. Тип вторичной структуры определялся окружением $x_1, x_2, \ldots, x_{n-1}, x_n$ из соседних аминокислот, расположенных слева и справа от исследуемой аминокислоты x_s (рис. 8). Вероятности последовательности $x_1, x_2, \ldots, x_{n-1}, x_n$ оценивались для моделей нестационарных цепей Маркова различных порядков по формулам вида (5), (9).

Таблица 1

Аминокислота	Частота		Пары аминокислот	Частота	
	1-я нить	2-я нить		1-я нить	2-я нить
A	0,08477	0,08648	AC	0,01153	0,01103
R	0,05099	0,05322	RV	0,06283	0,06112
D	0,05160	0,05491	DW	0,01883	0,01861
N	0,03633	0,03794	NE	0,04040	0,03936
C	0,01096	0,01116	CF	0,04485	0,04403
E	0,06098	0,06341	ET	0,06204	0,06181
Q	0,05099	0,05099	QD	0,04539	0,04593
G	0,06824	0,06961	GM	0,02125	0,02003
H	0,02227	0,02281	HA	0,06360	0,06597
I	0,06021	0,06129	IR	0,04989	0,05084
L	0,11011	0,11502	LN	0,03972	0,03978
K	0,04121	0,04225	KQ	0,05944	0,06075
M	0,02023	0,02045	MG	0,08411	0,08025
F	0,03890	0,04036	FH	0,02033	0,02148
P	0,05122	0,05266	PI	0,06176	0,06045
S	0,06500	0,06754	SK	0,03861	0,03956
T	0,05804	0,05933	TL	0,12268	0,12478
W	0,01533	0,01590	WS	0,06275	0,06673
Y	0,02803	0,02952	YP	0,05443	0,05434
V	0,06573	0,06640	VY	0,02683	0,02668

На выборке из 20 тысяч белков средний процент распознавания вторичной структуры белков на основе байесовских процедур распознавания на цепях Маркова составил 85 %.

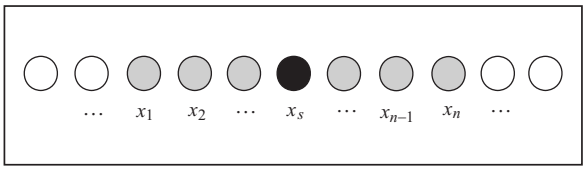

Рис. 8. Схема предсказания вторичной структуры одиночной аминокислоты

Заметим, что в литературе до сих пор не приведено объяснения феномена выполнения соотношений комплементарности в записи оснований по одной нити ДНК. Изложенные результаты показали, что эффективное применение байесовских процедур в процессе предсказания вторичной структуры белков, по сути, получено на основе выполнения соотношений комплементарности в геномах высших организмов. Соотношения комплементарности играют важнейшую роль в формировании пространственной структуры белковых молекул.

5. ФОЛДИНГ БЕЛКА

Фолдинг — процесс сворачивания полипептидной цепи в уникальную («нативную») пространственную структуру. Этот процесс всегда ведет к одной и той же пространственной структуре, для одной и той же цепи и длится менее секунды. Это наблюдение в свое время привело к гипотезе об уникальности пространственной структуры белка в зависимости от его аминокислотного кода.

Задачей фолдинга называется определение по аминокислотной последовательности белка его пространственной структуры, а именно, где расположены α-спирали, β-листы и участки *coil*; каким образом α-спирали, β-листы и участки *coil* образуют мотивы и домены.

Мотивом называют определенную последовательность элементов вторичной структуры белка. Как правило, это простая короткая последовательность, которая

встречается в нескольких белках. Например, спираль-*coil*-спираль. Этот мотив встречается во многих белках для связи с атомами кальция. Таким образом, у него есть вполне определенная функция.

Доменом называют более сложную, чем мотив, комбинацию вторичных структур с очень узкой функциональностью и имеющую активный центр, который может участвовать в связи с внешними молекулами. Доменов может быть один или несколько.

Существует множество подходов к решению задачи фолдинга, одним из которых является тредин (threading). Основная идея тредина заключается в том, что белки не сворачиваются в случайные структуры с бесконечным разнообразием вариантов. На самом деле, количество различных пространственных структур белков конечно, и можно даже выделить целый ряд определенных мотивов, присутствующих во многих белках. Так, например, только 15 % белков, добавленных в Protein Data Bank за последние несколько лет, можно считать обладающими новыми видами пространственной структуры. Все это позволило разработать метод тредина, состоящий в выравнивании белковой последовательности согласно той или иной предполагаемой пространственной структуре.

Существует множество алгоритмов тредина, но у них можно выделить общие черты. На начальном этапе тредина предполагается, что у исследователя есть для изучения аминокислотная последовательность белка с неизвестной пространственной структурой и база данных о белках, аминокислотная последовательность и пространственная структура которых известны (например, Protein Data Bank). На следующем шаге выполняется процедура сопоставления исследуемой цепочки с известными, последовательно, для всех возможных сдвигов цепочек относительно одна другой. При этом используется некая квазиэнергетическая функция, с помощью которой оценивают качество совпадения и выбирают одного или несколько лучших кандидатов. Затем на основе информации о пространственной структуре белков, выбранных на предыдущем шаге, строится некая последовательность вторичных структур, с заданным расположением в пространстве. На последнем шаге тредина исследуемый белок выравнивается по этой пространственной структуре [22, 23].

Приведем математическую формулировку выравнивания структуры к последовательности аминокислот (рис. 9), в которой определены:

- аминокислотная последовательность белка A, состоящая из n аминокислот $a_1 a_2 a_3 \ldots a_n$;
- оценочная функция выравнивания f; модель структуры белка C, состоящей из m вторичных структур, для каждой из которых известны длина c_i вторичной структуры i, а также то, что вторичные структуры i и $i+1$ соединены спиралью, для которой известны ее максимально и минимально возможные длины $l_i^{\max}$ и $l_i^{\min}$.

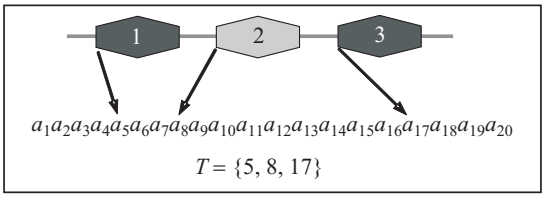

Рис. 9. Иллюстрация процесса выравнивания аминокислотной последовательности белка и его предполагаемой пространственной структуры

Для решения задачи необходимо найти такое множество $T = \{t_1, \ldots, t_m\}$, при котором значение f будет максимально. В множестве T t_i указывает, какая аминокислота из последовательности A является первой входящей в i-ю структуру.

Если игнорировать взаимодействие между парами аминокислот, то сформулированная выше задача решается методами динамического программирования, в противном случае она является *NP*-полной и поиск приближенного решения заметно усложняется.

Таким образом, полностью процедура тредина состоит из следующих шагов:

- выравнивание типа «цепочка»-«цепочка» и поиск наиболее подходящих кандидатов;
- построение предполагаемой пространственной структуры белка на основе пространственных структур найденных кандидатов;
- выравнивание типа «цепочка»-«структура».

Каждый из этих шагов ставит перед исследователями целый ряд сложных вопросов, что привело к появлению множества различных алгоритмов тредина. Тредин

дает хорошие результаты, но для действительно сложных задач используется в основном как вспомогательный метод для построения модели в первом приближении. Затем результаты трединга уточняются с помощью других методов и алгоритмов.

6. ОЦЕНКА КАЧЕСТВА

Существует два глобальных эксперимента по оценке качества предсказания пространственной структуры белков: CASP и EVA [24, 25].

CASP (Critical Assessment of Techniques for Protein Structure Prediction) — открытый для всех научных групп эксперимент, целью которого является изучение текущего состояния в области предсказания пространственной структуры белков, а также определение основных проблем и задач, критически важных для достижения успеха в этой области. В рамках CASP также проводится численная оценка качества предсказаний экспериментов каждой научной группы-участника, что превратило его в престижное соревнование. На данный момент в CASP принимает участие более 200 научных групп со всего мира.

В организационной структуре эксперимента можно выделить следующие группы:

1) организаторы — отвечают за все технические и организационные аспекты, связанные с проведением экспериментов и их обсуждением, начиная от выбора целей для прогнозирования и заканчивая организацией очных встреч участников эксперимента;

2) независимые оценивающие эксперты — группы по два человека, в каждой категории предсказаний, которые отвечают за оценивание качества предсказаний участниками и выявление основных существующих проблем в области прогнозирования белков; для оценки качества предсказаний эксперты пользуются утвержденными и согласованными ранее методами, но имеют право добавлять к ним собственные методики;

3) консультанты — группы, состоящие из предыдущих участников эксперимента (около десяти человек на каждую категорию предсказаний), влияющие на выбор методов независимыми оценивающими экспертами, а также на другие технические аспекты эксперимента;

4) организационное собрание участников — перед каждым очередным CASP происходит очное собрание его участников, на котором путем голосования решаются организационные и технические вопросы предстоящего эксперимента (дата проведения, выбор консультантов и экспертов), а также обсуждаются любые существенные изменения в самой процедуре эксперимента;

5) «Центр по предсказанию структуры белков» в Лаборатории Лоренса Ливермора — отвечает за сбор, управление и хранение всех данных эксперимента (данные по целям предсказаний, результаты предсказаний участников эксперимента, методы и результаты оценки предсказаний и т.п.).

CASP проводится раз в два года. В качестве целей для предсказания выбираются белки, третичная структура которых еще не известна, но будет исследована к окончанию эксперимента, либо известна, но нигде ранее не описывалась в открытом доступе. В CASP принимают участие как группы экспертов, так и полностью автоматизированные серверы. В первом случае машинные вычисления также используются, но не в таких объемах, как во втором; окончательная модель пространственной структуры исследуемого белка проверяется и исправляется человеком. Во втором случае вся работа по предсказанию и построению третичной структуры белка проводится компьютером. Следует отметить, что начиная с CASP-6 разница в итоговом результате между компьютерами и людьми очень незначительна, хотя преимущество все еще на стороне экспериментаторов.

Конкретные задачи, решаемые в рамках каждого эксперимента в CASP, следующие:

— предсказание третичной структуры белковых молекул (все CASP-эксперименты);

— предсказание вторичной структуры белковых молекул (отменено после CASP-5);

— предсказание белковых комплексов (только в рамках CASP-2, в настоящее время эта задача решается в рамках отдельного эксперимента CAPRI);

— предсказание биологической функции белка (начиная с CASP-6);

— предсказание контакта «аминокислота-аминокислота» в белке (начиная с CASP-4);

— оценка качества моделирования (начиная с CASP-7);

— распознавание границ доменов белка (начиная с CASP-6).

Задачи в рамках предсказания третичной структуры белков, в свою очередь, также делятся на две категории:

1) шаблонное моделирование (Template Modeling) — самый простой класс задач распознавания, к нему относят белковые молекулы, для которых существуют близкие родственные белки с известной третичной структурой;

2) нешаблонное моделирование (Template Free Modeling) — наиболее сложные для распознавания белки, не имеющие изученных ранее родственных аналогов.

В настоящее время отмечается прогресс качества CASP-экспериментов. Практически все участники эксперимента указывали на отсутствие значимых результатов начиная с CASP-5, т.е. с 2002 года [26]. Этот факт может свидетельствовать либо о недостатках существующих методов предсказания, либо об ограничениях со стороны вычислительных мощностей, доступных исследователям. Однако CASP-7 показал, что последнее не является основным ограничивающим фактором в предсказании третичной структуры белков, более того, группа, победившая в CASP-7, имела в своем распоряжении весьма скромные вычислительные мощности [27]. Таким образом, основная задача на данный момент состоит в усовершенствовании устаревших и разработке новых методик предсказания. Существуют отдельные категории задач, кроме распознавания третичной структуры белков, являющиеся подзадачами основной задачи, прогресс в решении которых очень важен для дальнейшего продвижения в этой области.

EVA — непрерывный во времени эксперимент, оценивающий качество предсказаний структур белков общедоступными серверами для следующих задач и методов:

— распознавание вторичной структуры белков;

— сравнительное моделирование (comparative modeling and homology modeling);

— метод трединга (protein threading).

В отличие от CASP в EVA не ставится никаких исследовательских задач. Основная цель эксперимента — постоянное информирование о качестве работы публичных серверов, предсказывающих структуры белков. В первую очередь этот проект важен для тех, кто не является экспертом в области предсказания структур белков, но использует информацию, полученную от общедоступных серверов, в своей работе или исследованиях. Проверка серверов-участников EVA производится в автоматическом режиме каждую неделю. В качестве целей для предсказания структур белков используются новые структуры, добавленные в Protein Data Bank в течение текущей недели.

EVA исследует качество работы серверов только для тех структур белков, которые подпадают под класс сложности шаблонного моделирования в CASP. Методы, использующие шаблонное моделирование в EVA, в настоящее время не оцениваются.

ЗАКЛЮЧЕНИЕ

В работе приведен обзор современных подходов предсказания пространственной структуры белков.

Наиболее перспективными, по мнению авторов, являются подходы, основанные на применении вероятностных моделей (модели цепей Маркова, условные случайные поля и т.п.). Модели строятся по информации из обучающих выборок, в качестве которых используются открытые банки данных белковых структур. Для заданной последовательности аминокислот, или наблюдений, требуется найти наиболее вероятную последовательность состояний. Состояниями могут быть типы вторичной структуры или торсионные углы в зависимости от конкретной задачи. Так, например, байесовские процедуры на цепях Маркова различных порядков довольно успешно предсказывают вторичную структуру белков.

Особенность бактериальных геномов позволяет выделять и отдельно анализировать аминокислотные последовательности белок-кодирующих участков. Получены новые

важные данные о совпадении оценок переходных вероятностей аминокислотных последовательностей белков, синтезированных на двух противоположных нитях ДНК бактерий. Это нельзя объяснить случайным совпадением, поскольку при подсчете оценок переходных вероятностей фигурируют 399 независимых параметров. Полученный вывод подтвержден численными расчетами на геномах бактерий. Данный результат, по сути, позволил подтвердить не очевидную до этого эффективность использования байесовских процедур распознавания на цепях Маркова для предсказания вторичной структуры белков.

СПИСОК ЛИТЕРАТУРЫ

1. Ginalski K., Grishin N.V., Godzik A., Rychlewski L. Practical lessons from protein structure prediction // Nucleic Acids Res. — 2005. — **33.** — P. 1874–1891.
2. Lazaridis T., Karplus M. Effective energy functions for protein structure prediction // Current Opinion in Structural Biology. — 2000. — **10.** — P. 139–245.
3. Boas F., Harbury P. Potential energy functions for protein design // Ibid. — 2007. — **17.** — P. 199–204.
4. Narang P., Bhushan K., Bose S., Jayaram B. A computational pathway for bracketing native-like structures for small alpha helical globular proteins // Phys. Chem. Chem. Phys. — 2005. — **7**. — P. 2364–2375.
5. Madhu Smitha, Abhijit Mitra, Harjinder Singh. Real valued genetic algorithm based approach for protein structure prediction — role of biophysical filters for reduction of conformational search space // Third IAPR Intern. Conf. on Pattern Recognition in Bioinformatics PRIB, Oct. 15–17 2008, Novotel St Kilda. — Melbourne, Australia, 2008.
6. Reconstruction of 3D Structures from protein contact maps / M. Vassura, L. Margara, P. Di Lena et al. // IEEE/ACM Trans. on Comput. Biology and Bioinformatics. — 2008. — **5(3).** — P. 357–367.
7. Pollastri G., Baldi P. Prediction of contact maps by GIOHMMs and recurrent neural networks using lateral propagation from all four cardinal corners // Bioinformatics. — 2002. — **18.** — P. 62–70.
8. Cheng J., Baldi P. Improved residue contact prediction using support vector machines and a large feature set // BMC Bioinformatics. — 2007. — **8.** — P. 1–9.
9. Сергиенко И.В., Гупал А.М. Статистический анализ генома // Цитология и генетика. — 2004. — № 4. — С. 76–81.
10. Viterbi A.J. Error bounds for convolutional codes and an asymptotically optimal decoding algorithm // IEEE Trans. Informat. Theory. — 1967. — **IT-13.** — P. 260–269.
11. Baum L. E., Petrie T. Statistical inference for probabilistic functions of finite state Markov chains // Ann. Math. Statist. — 1966. — **37.** — P. 1554–1563.
12. Levinson S.E., Rabiner L.R., Sondhi M.M. An introduction to the application of the theory of probabilistic functions of a Markov process to automatic speech recognition // Bell Syst. Tech. J. — 1983. — **62.** — P. 1035–1074.
13. Lafferty J.D., Mc Callum A., Pereira F.C.N. Conditional random fields: probabilistic models for segmenting and labelingsequence data // Proc. of the Eighteenth Intern. Conf. on Machine Learning (ICML 2001). — New York: Morgan Kaufmann Publ., 2001. — P. 282–289.
14. The international human genom sequencing consortium // Nature. — 2004. — **431.** — P. 931–945.
15. Гупал А.М., Сергиенко И.В. Оптимальные процедуры распознавания. — Киев: Наук. думка, 2008. — 232 с.
16. Гупал А. М., Вагис А.А. Комплементарность оснований в хромосомах ДНК // Проблемы управления и информатики. — 2005. — № 5. — С. 153–157.
17. Сергиенко И.В., Гупал А.М., Вагис А.А. Соотношения комплементарности в записи оснований по одной нити ДНК // Цитология и генетика. — 2005. — № 6. — С. 71–75.
18. Anderson T.W., Goodman L.A. Statistical inference about Markov chains // Ann. Math. Statistics. — 1957. — **28.** — P. 89–110.
19. Белецкий Б.А., Васильев С.В., Гупал А.М. Предсказание вторичной структуры белков на основе байесовских процедур распознавания // Проблемы управления и информатики. — 2007. — № 1. — С. 61–69.
20. Сергиенко И.В., Белецкий Б.А., Васильев С.В., Гупал А.М. Предсказание вторичной структуры белков на основе байесовских процедур распознавания на цепях Маркова // Кибернетика и системный анализ. — 2007. — № 2. — С. 59–64.
21. Белецкий Б.А., Вагис А.А., Васильев С.В., Гупал А.М. Процедуры распознавания вторичной структуры белков // Проблемы управления и информатики. — 2007. — № 4. — С. 134–139.
22. Thorda A. Protein threading. — Hambourg: Univ. of Hambourg, 2003. (http://en.scientificcommons.org/40891925)
23. Mc Guffin L.J. Protein fold recognition and threading in computational structural biology // World Scientific. — 2008. — P. 37–60.
24. http://predictioncenter.org/
25. http://cubic.bioc.columbia.edu/eva/
26. Moult J., Krzysztof F., Zemla A., Hubbard T. Critical assessment of methods of protein structure prediction (CASP) — Round V // Proteins. — 2003. — **53.** — P. 334–339.
27. Kryshtafovych A., Krzysztof F., Moult J. Progress from CASP6 to CASP7 // Ibid. — 2007. — **69.** — P. 194–207.

Поступила 02.06.2009